Fact Sheet

National Center for Biotechnology Information ■ National Library of Medicine ■ 8600 Rockville Pike ■ Bethesda, MD 20894

NCBI's Web Services

Established in 1988 as a national resource for molecular biology information, the National Center for Biotechnology Information (NCBI) creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedial information. Through its web services at *www.ncbi.nlm.nih.gov*, NCBI provides integrated access to the GenBank® DNA sequence database, the human genome, more than 40 related molecular biology database services, and the scientific literature. Selected resources are highlighted here.

Resource Highlights

Databases

■ **Entrez** provides integrated access to nucleotide and protein sequence data from GenBank, RefSeq, and several protein databases, along with 3D protein structures, gene mapping and phenotype information, the NCBI taxonomy, and related journal articles in PubMed. Entrez contains precomputed similarity searches for each database record, producing a list of related sequences, structures, and journal articles.

■ **The Entrez Genomes** database presents graphical displays of entire genomes and chromosomes, using sequence maps integrated with genetic and physical maps. Genomes can be searched using an extensive list of gene markers and lists of the proteins encoded within genomes are also available.

■ **The Human Genome.** A challenge facing researchers today is how to piece together and analyze the multitudes of data currently being generated through the Human Genome Project. NCBI has an ongoing program of incorporating new data and annotation into its suite of human genome resources and producing updated assemblies on a regular basis. The *Human Genome Map Viewer* provides integrated access to genome data through more than 20 genetic, physical, and sequence maps, with views ranging from specific genes to whole genomic regions of interest.

■ **LocusLink** provides a single query interface to curated information about genetic loci, including nomenclature, phenotypes, DNA and protein sequences, EC numbers, UniGene clusters, homology, and map locations. Links are provided to NCBI resources plus a diverse array of related resources.

■ **Specialized databases for EST and STS data** include the dbEST database of expressed sequence tags, the dbSTS database sequence tagged sites, and the UniGene database, which contains more than 97,000 human sequence clusters that represent the transcription products of distinct genes.

■ **The Online Mendelian Inheritance in Man** (OMIM) database is a continuously updated catalog of human genes and genetic disorders, and is considered a phenotypic companion to the human genome project. Each database entry contains a comprehensive state-of-the art review, plus links to associated records from GenBank, MEDLINE, and the OMIM gene map.

■ **The Molecular Modeling Database** (MMDB) presents macromolecular 3D protein structure data in a way that integrates chemical, sequence, and structure information. Structures have been compared using VAST (Vector Alignment Search Tool) to identify significantly similar 3D substructures. Links to NCBI's Conserved Domain Database (CDD) and the Domain Architecture Retrieval Tool (DART) are also provided. A 3D structure viewer, Cn3D, may be installed in your Web browser for interactive visualization of MMDB structures and structural neighbors.

■ **The Taxonomy** browser allows users to retrieve nucleotide and protein sequences for a particular taxon, and to browse up and down the taxonomic tree. The NCBI taxonomy contains the scientific names, common names, and synonyms of all organisms represented in the sequence databases.

Databases (continued)

- **PubMed** provides access, free of charge, to MEDLINE, a database of more than 10 million bibliographic citations and abstracts in the biomedical journals. Links are provided to full-text articles at participating publishers' Web sites and to NCBI's molecular biology databases.

Sequence Analysis Tools

- **BLAST** (Basic Local Alignment Search Tool) sequence similarity search services are available for comparing a nucleotide or protein sequence against a collection of databases. Database selections include non-redundant nucleotide and protein databases, a rolling month database for data added within the last 30 days, dbEST, dbSTS, the human genome, and other specialized databases.

- **Electronic-PCR** makes it possible to determine the gene map location of a new sequence if it contains an STS. *ORF Finder* is a graphical analysis tool which finds all open reading frames in a sequence.

- **Clusters of Orthologous Groups** (COGS) were delineated by comparing protein sequences encoded in seven complete genomes representing five major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least three lineages, and thus corresponds to an ancient conserved domain. A tool for comparing your sequence to the COG database is available.

DNA Sequence Submission to GenBank

- **BankIt** is a Web sequence submission tool that uses a simple forms-based approach to formatting and submitting a sequence to GenBank. Accession numbers are assigned within 24 hours. New submissions become GenBank records after extensive quality assurance processing by NCBI staff and consultation, if necessary, with the author.

- **Sequin**, a stand-alone software tool for submitting DNA sequences to GenBank, can be downloaded from the NCBI Web site. Sequin is designed to simplify multiple sequence submissions, provide graphical viewing and editing options, and accommodate long sequences and complex submissions. Custom protocols for high volume submissions such as ESTs, STSs, and complete genomes are also available and documented on the GenBank Web page.

A challenge facing researchers today is the ability to piece together and analyze the multitudes of data generated through the Human Genome Project. NCBI's Web site serves as an integrated, one-stop, genomic information resource for biomedical researchers around the world.

FTP Service

- **NCBI's FTP site** provides access to a collection of software and databases. Client programs for Entrez and Network BLAST are in the *entrez* and *blast* directories. The Sequin data submission software is in the *sequin* directory. The most current GenBank release and daily updates are in the *genbank* directory. The repository directory contains a collection of more than 40 contributed molecular biology databases, maintained and updated by the contributing curators. The NCBI toolbox, in the *toolbox* directory, contains a set of public domain software tools and data exchange specifications that developers may use to produce portable, modular software for molecular biology.

Research in Computational Biology

NCBI's research program focuses on theoretical, analytical and applied approaches to a broad range of fundamental problems in molecular biology. From the Home page, click on *Research at NCBI* for descriptions of research projects, a staff bibliography, and the full text of selected publications.

National Center for Biotechnology Information ▪ 8600 Rockville Pike, Building 38A, Rm 8N-803
Bethesda, MD 20894 Phone: 301-496-2475 ▪ Fax: 301-480-9241 ▪ e-mail: info@ncbi.nlm.nih.gov
web site:
www.ncbi.nlm.nih.gov
revised 02/2002